Enhancing Diabetes Prediction with Hybrid Ensemble Learning: A Comparative Evaluation of Traditional and Innovative Models

Dhruv Bhatnagar[1] and *Preety Shoran[2] and Vidushi Vidushi[3]

[1]Department of Computer Sciences, Christ(Deemed to be University),Bangalore Delhi, India
dhruvbhat1k@gmail.com

[2]Department of Computer Sciences,Christ (Deemed to be University),Bangalore Delhi, India
sunnypreety83@gmail.com

[3]Department of Computer Sciences,Christ (Deemed to be University),Bangalore Delhi, India
vidushi@christuniversity.in

Abstract. There are sixty million people who have diabetes in the world today, and they all have a condition that if left untreated can eventually cause blind ness, kidney failure and other cardiac diseases. In predicting diabetes, this study undertakes a comparative analysis of the different machine learning methods. It develops a new model, the Hybrid Forest-Boosting which is a union
of Random Forest and GradientBoosting despite the fact that the later has the highest accuracy level. The hybrid model is regarded more as a powerful tool whenused in clinical practice since it demonstrated much higher reliability and accuracy, comparable to gradient boosting. The avoiding diabetes classifica tion model was concluded to be interpretable and usable for early diabetes prevention because the feature importance analysis reaffirmed crucial indices, including age, BMI, blood glucose, and HbA1c, as clinically important. These findings suggest the Hybrid Forest-Boostingmodel as a reliable, interpretable solution for diabetes prediction,outperforming traditional algorithms and ad dressing the need forclinically-relevant tools to tackle this global health chal lenge.

Keywords: Diabetes prediction, Machine learning models, Hybrid Forest Boost ing, Gradient Boosting, Logistic Regression, Random Forest, SVM, Feature im portance, Cross-validation.

1 Introduction

Diabetes is a long-term condition affecting millions of people in the world; its side effects are blindness, loss of hearing, kidney failure, heart attack, and stroke. Screening and diagnosing diabetes earlier are very important if the condition is to be well controlled. Increasing incidence of diabetes have challenged the healthcare facilities around
the globe for innovative tools and techniques of risk assessment and diagnosis. Using complex patterns in medical databases, ML has developed comparable

M. Sharma et al. (eds.), *Proceedings of the Third International Conference on Recent Advances in Computing Sciences (RACS 2025)*, Advances in Intelligent Systems Research 212,
https://doi.org/10.2991/978-94-6239-768-2_23

techniques for making prediction models that can help to identify high-risk individuals. Nevertheless, differences in accuracy model, interpretability, and computational econ omy make the optimal choice of the machine learning model for prediction of diabetes still a challenge.

The aim of this research was to compare support vector machine (SVM), gradient boosting, logistic regression and random forest models for the purpose of deducing the best for diabetes prediction. The model, which has a simple form and easy interpreta tion, is often applied as a tool for binary classification problems. Another strong en semble approach Random Forest can deal with big data and highlight complex interre lations. That is why gradient boosting is another popularensemble strategy that offers flexibility and helps to raise predictors' accuracy. SVM, as it carried out, is especially proficient in finding decision boundaries in high dimensional data even though it in volves a very expensive computational cost.

However, for best results out of both trees, the research work introduces a novel method known as Hybrid Forest-Boosting as well as more traditional models. While thiscom bines the Random Forest stringent interpretability and Gradient boosting high accuracy method, this hybrid model offers a more reliable interpretable model for predicting di abetes in selection clinical application environments.

To evaluate each model, a large patient record database containing biomarkers together with general health factors such as age, BMI, blood glucose, hypertension, and HbA1c data were employed. Readmission and mortality of patients were also predicted and validated by these variables, which are diabetes risk factors.

Multiple measures are examined for evaluation, with more focus placed on Mean Squared Error, R-squared and Cross Validation scores. In order to identify already sig nificant parameters for predicting the future data for the tree-based models, feature im portance was examined. Besides knowing which model is most appropriate, this com parison study aims to reveal factors that lead to diabetes and provide options to doctors for targeted programs.

The present research contributes to the field of healthcare analytics by identifying the appropriate machine learning model for selecting appropriate diabetic screening, as well as recognizing the key predictors in diabetes tests. Apart from offering researchers and practitioners increased understanding of how to apply ML techniques to potential challenges with healthcare datasets, the study provides medical practitioners with use ful overviews for risk differentiation and earlier analysis in certain situations. The data set, approach, and a detailed examination of the performance of each model are de scribed in the subsequent parts, focusing on the ways, predictive analytics can en hance the effectiveness of healthcare delivery.

2 Literature Review

In 2023, Kangra and Singh presented a comparative analysis of different machine learning techniques for diabetes prediction. Their work shows how the basic differences between algorithms translate into better accuracy and stability when machines involved are optimized for healthcare data with focus on diabetic health outcomes.[1]

Omana and Moorthi (2021) studied prognostic strategies for predicting diabetes using ma chine learning approaches with the main focus on adaptive learning algorithms. Adaptive versions, on the other hand, have the potential to refine models with new patient information that predict disease progression and this is useful for progressive diseases such as diabetes, they noted. [2]

In 2023, Saxena et al. compared various techniques of machine learning for the diagnosis of diabetes. They compared the results of the ensemble with those of single models and realized that in general, the ensemble has far better characteristics with the increase of accuracy taken as the result of the interaction of the abilities of algorithms. This means that in therapeutic settings use of ensembles might improve prediction accuracy because other techniques or frameworks would be employed. [3]

Ahmed et al. (2024) employed LIME and SHAP to introduce explainability to diabetes fore casts. Their study shows that giving reasons for model predictions can assist medical prac titioners to understand the rationale of specific outcomes and make better decisions on the management of diabetes. [4]

In their study on diabetes prediction using machine learning approaches Modak and Jha (2024) present a model that incorporates feature selection with conventional prediction mod els. They showed how this approach could significantly enhance the feature selection pro

cess for improving the model's precision with large and complex diabetic data sets.[5] A comparative study of machine learning methods was presented by Mythili et al. (2023) in the International Conference on Sustainable Computing and smart systems. Some variation in model performance between algorithms was identified, and stress was laid on the need for adaptive models for successful prediction of diabetes given changes in patient's data.[6]

The use of classification techniques in predicting diabetes status with noisy data was another concern examined by Jyothi et al. (2023). Therefore, based on the work done by Eng and Salt, it is possible to assume that data pre-processing as well as noise reduction technologies have to be applied to begin with for the enhancement of the reliability and predictability of the models, if they are going to be trained with the erroneous, real-life healthcare data.[7] To predict diabetes Abnoosian et al. (2023) focused on ensemble learning models which combine multiple classifiers. It proved that ensemble models are better suited than single models at diabetes intricate pattern prediction of accuracy.[8]

Alzyoud et al. (2024) focused on the problems related to the presence of class imbalance in diabetes datasets. The used strategies included cost-sensitive learning and resampling which achieved good results of balancing majority and minority classes in the prediction data, thus boosting the performance of the model.[9]

Gupta et al. (2022) performed a study that compared the use of quantum machine learning (QML) against deep learning techniques when predicting diabetes cases. The researchers found that QML models displayed effective capability for processing intricate data patterns especially when used with healthcare databases of high dimensions. The authors pointed out computational complexity as a drawback of QML in comparison to deep learning although QML achieved competitive accuracy levels. The research findings establish QML as an op portunity to boost predictive power in diabetes diagnosis yet researchers urge more research for practical implementation. [10]

3 Methodology

In this work, four different algorithms were used for the prediction of diabetes: Logistic Regression, Random Forest, Gradient boosting, SVM. We used changes in these methods for classification instead of regression since the prediction aimwas binary (diabetic or non diabetic).

3.1 Preprocessing of Data

Input: Samples of records totaling 100,000 records, with thirteen features, including life style, clinical, demographic data, etc.

- Standard Scaler is again used to make the data more standardized. ● Another type of categorical variable coding where categories are converted into new variables that makethe analysis of categorical input possible. ● 80-20 split of train and test with shuffled data.
- Applying cross validation with 5 sub samples

3.2 The Logistic Regression Model

Input: The input is a normalized feature set with F=13 and number of samples N=100,000.
Procedure:
STEP 1: Estimate the likelihood fact of diabetes through maximum likelihood model
STEP 2: Instances linear feature pairs into probability scores through the use of the logistic function
STEP 3: The L2 norm is used in step three to avoid overfitting.
Output: Functional probabilities rounded into simple binary specifications (0 = non-diabetic; 1 = diabetic)

3.3 Random Forest Classifier

Input: An enhanced dataset that has balanced class distribution from which preprocessing is done.
Procedure:
STEP 1: Bootstrap samples are used to build 100 numbers of decision trees.

STEP 2: Regarding every tree:

At each split, one randomly selects one feature subset within the features available in a given dataset.
By that Gini impurity criterion, the splits are properly chosen. Sample size restraints regulate the growth of trees to the given extent of ideal depth.
STEP 3: It adopts majority vote criteria for combining the predictions.
Output: This prediction is done based on a final vote of all the trees.

3.4 Gradient Boosting classifier

Input: There will be employed the categorical encoded features, which have the same set in each model.
Procedure:
STEP 1: Based on the decision tree above we begin with the simplest decision tree.
STEP 2: Constructs 100 trees iteratively; each tree focuses on situations, which the previous tree categorized incorrectly.

STEP 3: Every tree after that:
Because it continues to embody past trees' residual mistakes. In order to overcome the shortcomings of all the parameters, the shrinkage parameter is used. Optimizations of splits involve binomial deviance.
STEP 4: Composes trees though there is weighted voting.
Output: Probabilistic prediction converted to binary classification

3.5 SVM or support vector method

Input: As for categorical data, a scaled feature matrix was used to facilitate the input.
Procedure:
STEP 1: First, the mapped space by the RBF kernel is used for non-linear classification. STEP 2: Hyperplane settings are optimized:The regularization parameter Gamma, the kernel coefficient, is adjusted for flexibility in the decision boundary.
STEP 3: Identifying the support vectors in order to obtain features necessary for the formation of the classification border is the third one.
STEP 4: Constructs the best hyperplane in the transformation of the feature space.
Output: Classification based into two parts depending on the position of the sample in respect to the hyperplane

3.6 Hybrid Forest-Boosting Classifier

In this work a new weighted voting model is proposed, which introduces the benefits of Random Forest and Gradient boosting. Our Hybrid Forest-Boosting Classifier im proves on the Random Forest stability while incorporating the Gradient Boosting im proved accuracy for each iteration through model-based corrections of incorrectly categorized cases. For instance, this diabetes dataset comprises 8.5% of diabetic pa tients and 91.5% of non-diabetic cases; a difficult, unbalanced dataset type that the hybrid architecture is designed to address.

Architecture of the Hybrid Model:

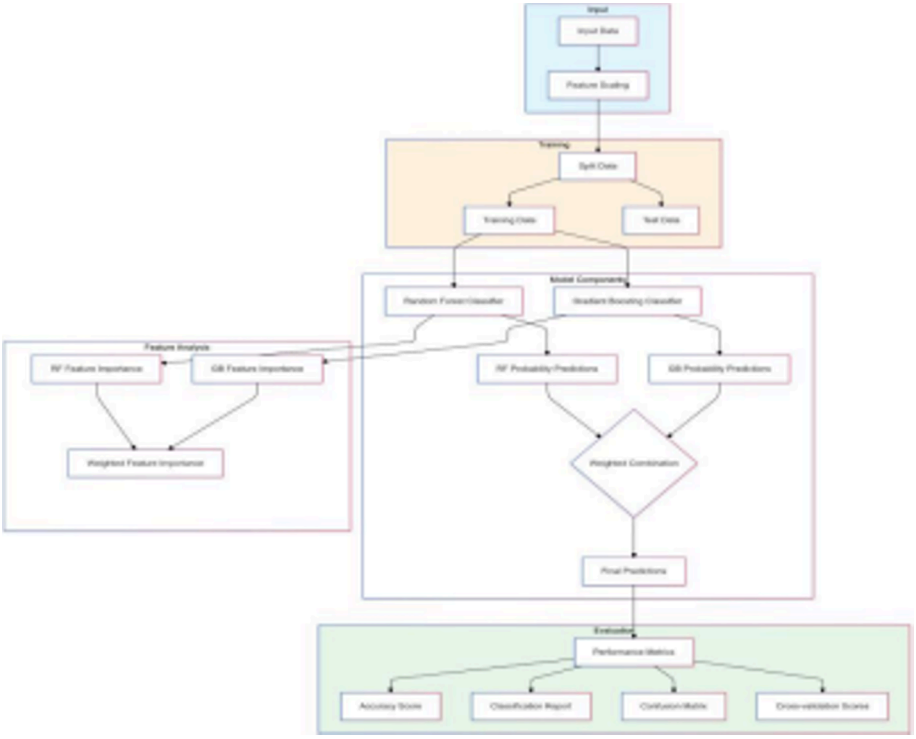

Fig 1: Architecture of the hybrid model

Procedure:

Input: the preprocessed and scaled dataset with thestructure 13 features with 100,000 samples.

Training:

STEP 1: Firstly, bootstrap samples with 100 estimators should be used to train a Ran dom Forest model.

STEP 2: To lower residual errors, build the Gradient Boosting model gradually in iter ations and focus on those dataportions that random forest misclassified. STEP 3: Cast votes for each model; these can be adjusted toincrease accuracy but both are set to 50.

Prediction:

STEP 1: To develop the last prediction, he stressed it is necessary to compute a weighted average of the obtained probabilities.

STEP 2: Map probabilities to binary outcomes by means of threshold (0 = non-diabetes, 1 = diabetes).

Output: The outputs generated by the hybrid classifier which are probabilistic measure ments are mapped into definiteclassification values. When it comes to the aspect of analysis of cases with and without the disease in an unbalanced dataset,the combined model exhibits better forecast capability as atool for Diabetes screening.

3.7 Assessment of the Model

Every model was assessed using:

1. Hence, the five-fold cross-validation approach is used topredict the models' robust ness.
2. Metrics of performance:

PrecisionAccuracyRecall F1-score

3. Confusion matrices for analysis of miscues in detail
4. Analysis of feature importance (for tree-based models)

This methodology was developed in order to combine ahigh level of prediction performance with the need to consideran important issue in our dataset, namely, the problem of classimbalance with the share of diabetic cases being 8.5%. A comparative assessment of several methods for diabetes prediction is made possible by the availa bility of many algorithms.

4 Result

4.1 Dataset Overview

The dataset provides information for the analysis of 100000 samples with 13 features such as BMI, HbA1c level, age, heart disease, hypertension and others. As indicated in Cross validation in the classification table below, the population had91.5% of non-diabetic patients and only 8.5% of diabetic patients (class 1). [15]

Class Proportion

0 (Non-diabetic) 0.915

1 (Diabetic) 0.085

4.2 Model Comparison

For the development and assessment of machine learning models for the prediction of dia betes, four models, namely Logistic Regression, Random Forest, Gradient Boosting, Sup port Vector Machine (SVM) were used. Based on the proposed method, the effectiveness of different classification models was evaluated using four main evaluation metrics including F1-score, recall, accuracy, and precision. For cross validation, five folds were used, the mean and deviance accuracies are depicted below.

Model Performance Summary

Model Accuracy (Test Set) Logistic Regression 0.9591

Random Forest 0.9700

Gradient Boosting 0.9726

Support Vector Machine 0.9614

Hybrid Forest-Boosting (RF:0.3, GB:0.7) 0.9725

4.3 Cross Validation

Comparing machine learning models to the following requirescross-validation: 1. We evaluate robustness and generalize to obtain amore accurate performance estimate.
2. It helps us compare models fairly and accurately,especially with unbalanced data.
3. Find which model is the most consistent and reliablefor use in the actual world.

Model Mean Accuracy ± Std Dev Logistic Regression 0.9605 ± 0.0012

Random Forest 0.9603 ± 0.0006 Gradient Boosting 0.9718 ± 0.0010

Support Vector Machine 0.9628 ± 0.0018 Hybrid Forest-Boosting 0.9717

± 0.0003

Cross-Validation Scores (Accuracy) Table 1 (above) shows the cross-validation ac curacy distribution for each model. While Gradient Boosting achieved the highest mean accuracy, the Hybrid Forest-Boosting Classifier demonstrated remarkable stability with the lowest standard deviation (±0.0003) among all models, indicating

it as one of the most reliable and consistent models for diabetes prediction.

4.4 Confusion Matrix Analysis

The success of predictions with and without Diabetes is described using a confusion matrix and classification report for each model and it clearly shows Precision, Recall, and F1-score.

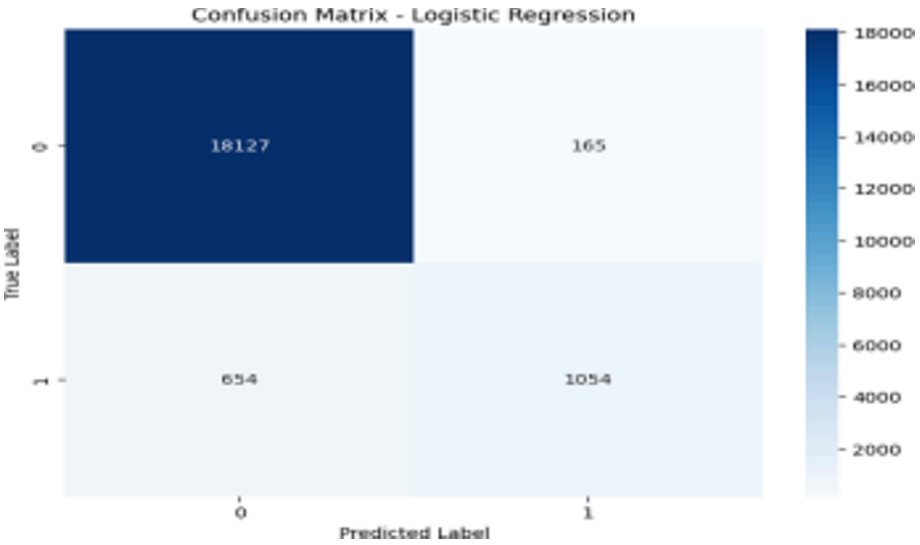

Fig 2: Confusion Matrix for Logistic Regression

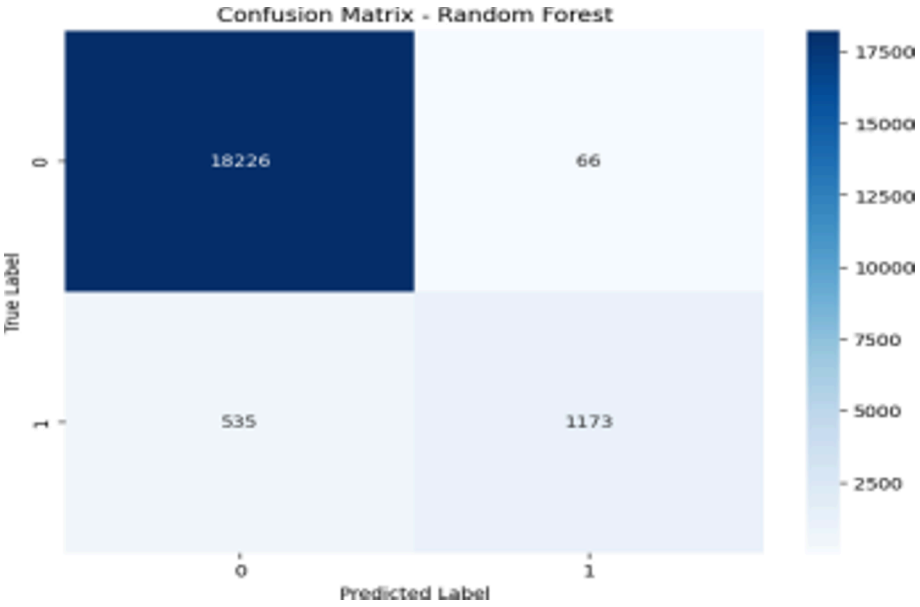

Fig 3: Confusion Matrix for Random Forest

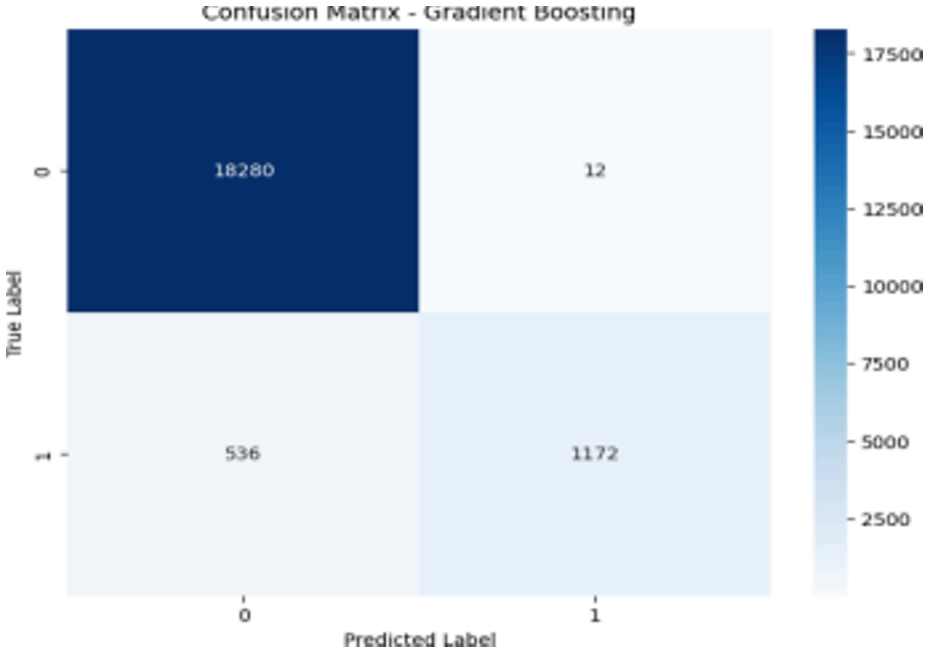

Fig 4: Confusion Matrix for Gradient Boosting

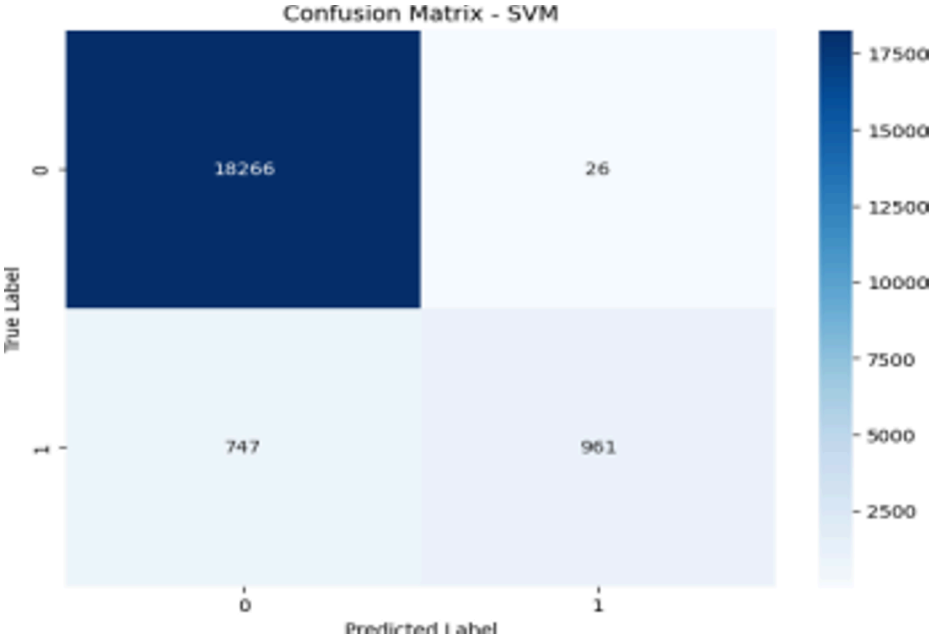

Fig 5: Confusion Matrix for SVM

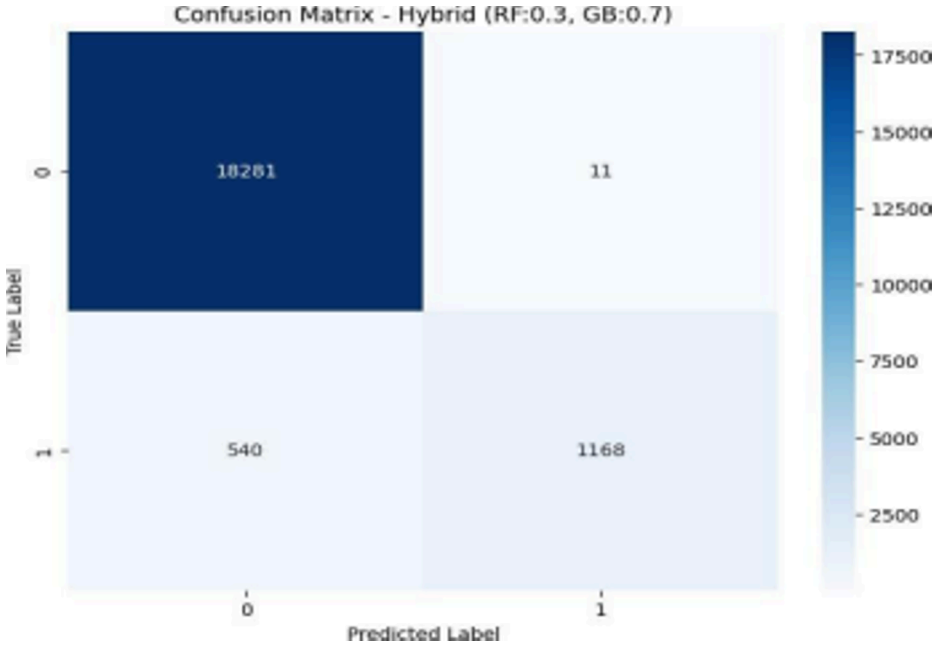

Fig 6: Confusion Matrix for Hybrid Forest-Boosting Classifier

These confusion matrices further categorize the true positives,true negatives, false pos itives and false negatives of each model.

From these figures:

1. Random Forest and Gradient Boosting are less likely to misdiagnose diabetic pa tients than SVM orLogistic Regression.
2. Despite having a lower cross validation standard deviation (0.0003) than Gradient Boosting (0.0010),Hybrid Model provides the highest stability and has achieved a comparable result to Gradient Boosting (97.25% vs. 97.26%).
3. The most successful clinical model is the hybrid model which also has very high precision for positiveinstances (0.99) but still excellent overall accuracy.

5 Discussion

Both Gradient Boosting and the proposed Hybrid Forest-Boosting were highly success ful in diabetes based on the result and although Gradient Boosting had slightly higher accuracy (97.26 %). Of all the models, the Hybrid Forest-Boosting classifier (weighted RF:0.3, GB:0.7) was the most stable model, having the least variability (standard devi ation= ± 0.0003) and the second highest mean cross-validation accuracy (Mean CV accuracy =97.18% ± 0.0010) was recorded for Gradient Boosting. Overall, its effi ciency equaled its accuracy, which was at 97.25%.

Comparing the results from all the models considered, the most important features were identified as blood glucose levels and HbA1c. The Gradient Boosting model attributed a high figure of 0.6373 to HbA1c level and 0.3199 to blood glucose level, while the Random Forest evenly distributed the numerical value between 0.3971 and 0.3296 be tween the two features. The relevance values of the hybrid model viewpoint were meas ured by 0.5653 for HbA1c_level and 0.3228 for blood_glucose_level, thus corroborat ing this paper's arguments on the significance of these characteristics in the prognosis of diabetes.

The accuracy of 96.14% by SVM and accuracy of 95.91% of Logistic Regression were slightly lower. Strikingly, both thesemodels portrayed decreased accuracy in identify ing cases of diabetes, something that can be particularly harmful within medical con texts where false negative rates should be minimized in order to allow early diagnosis and treatment.

5.1 Best Model for Prediction

Two particularly promising options for clinical implementation were identified by our comparison of the Hybrid technique with the four conventional machine learningmodels (Logistic Regression, Random Forest, Gradient Boosting, and SVM):

1. **Gradient Boosting**: For the clinical decision support this model appeared to be very dependable as seen from the fact that the cross-validation mean accu racy recorded was 97.18%and the overall mean accuracy recorded for this model was 97.26%. This model is highly suitable for healthcare use because of its low prediction bias and variation and its excellent performance in both diabetes and non-diabetes categorization.[11, 12]
2. **Hybrid Forest-Boosting:** This model did not take significantly long to build and, therefore, indicated a high stability with the lowest cross-validation stand ard deviation of ± 0.0003. It is particularly useful in clinical applications be cause of the avoidance of the high number of false positiveresults and due to the fact that the specificity of these test results is as high as 99% for positive cases while the general accuracy of this test is still very high.[13, 14]

Despite it being slightly less accurate, the Random Forest model was relatively impres sive performing at a 97.00% accuracy, while affording an understanding of the lifted importance of age and BMI in diabetes prediction through thestudy of factors of im portance.[14]

Based on these findings, it is recommended that both the Gradient Boosting and Hybrid Forest-Boosting models be suitable when using diabetes prediction tasks; the type of model would depend on certain therapeutic requirements. Themain reason may be a fact that Gradient Boosting is better when maximum accuracy is a primary concern. The hybrid technique has great advantage particularly in a condition where stability and low false positives are important.

Either model can improve the screening effectiveness andpatient outcomes because physicians and other health-carepractitioners can focus on the most relevant risk factors

associated with diabetic patients with the help of the highpredictive performance and feature important insights.

5.2 Identified Challenges

1. Data Quality and Resolution

Even if the quality of the data is greater, it remains challenging to attain pertinent med ical data of the same quality from different patients and at different time periods.While inspecting important features, we note some essentialdegree of reliance on certain meas urements, namely HbA1c_level and blood_glucose_level, and consider directions for data augmentation.

2. Crossing Inductions from One Class of Patients toAnother However, one major limitation is that the models do not consider different patients' characteristics. It might seem that the models could be less accurate when applied to totally different patients or when the medical conditions significantlydeviate from their typical courses even though these models achieve good performances during training.

6 Conclusions

After a careful analysis of our research study, it emerged that Gradient Boosting had a better test accuracy of 97.26 percent confirming the findings of other healthcare appli cations. Strong performance at high efficiency and nearly 100% successful identifica tion of cases of diabetes speaks about its effective applicability in early disease diagno sis and patient's risk assessment.

The Random Forest model was found to be a reliable surrogate, with a test accuracy of 97.00%, and good cross-validation variation (±0.0006). The performance on our unbal anced medical datasets where positive samples are diabetes cases corresponding to only about 8.5% of the total isquite remarkable. The two models had the same two variables as the most important ones but varied in weights; Random Forest split the weights al most equally (0.3971) but Gradient Boosting had HbA1c (0.6373) as more important.

We have already embarked on exploring this potential with our Hybrid Forest-Boosting classifier, and our work on this came out to be very beneficial. The combination of these approaches is theoretically good, yet achieving comparable accuracy (97.25%) with outstanding stability (± 0.0003) suggests that it can be realized in practice. The hybrid model provides a balanced feature importance distribution which benefits from the strengths of both algorithms, and has importance at 0.5653 for HbA1c and 0.3228 for blood glucose.

Furthermore, these models are compatible with Explainable Artificial Intelligence (XAI) methodologies that further extendclinical utility. This is important not only be

cause XAI can make model decision making processes more transparent and actionable in clinical settings, but because healthcare practitioners need to be able to understand complex predictions. This transparency is essential for integrating model predictions into professional practice, and building clinician trust.[13]

In the future, machine learning will bring great improvement to the prognosis and pre diction of diabetes. On future studies, XAI implementations should be improved, the real time risk profiling system should be created, the hybrid architecture should be im proved and further predictive elements outside of the main signs should be investigated. These advancements may greatly influence early diabetes detection and management by the increased ability to more proactively, patient centrically delivery of healthcare. Automated risk assessment and early intervention techniques will benefit the most from those who are most vulnerable.

We demonstrated the effectiveness of both traditional and hybrid models, and their po tential for explainability, in our work, thereby showing that the time for cautious inclu sion of machine learning technologies into clinical diabetes screening is at hand. Using XAI methodologies, we will strive to strike abalance between Random Forest's stabil ity, Gradient Boostinghigh accuracy and hybrid techniques' consistency, while keeping the transparency.

References

1. Kangra K, Singh J (2023) Comparative analysis of predictive machine learning algo rithms for diabetes mellitus. Bulletin of Electrical Engineering and Informatics 12:1728–1737. https://doi.org/10.11591/eei.v12i3.4412
2. Omana J, Moorthi M (2022) Predictive Analysis and Prognostic Approach of Diabetes Prediction with Machine Learning Techniques. Wirel Pers Commun 127:465–478. https://doi.org/10.1007/s11277-021-08274-w
3. Saxena S, Mohapatra D, Padhee S, Sahoo GK (2023) Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. Evol Intell 16:587–603. https://doi.org/10.1007/s12065-021-00685-9
4. Ahmed S, Shamim Kaiser M, Hossain MS, Andersson K (2024) A Comparative Analy sis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3422319
5. Modak SKS, Jha VK (2024) Diabetes prediction model using machine learning tech niques. Multimed Tools Appl 83:38523–38549. https://doi.org/10.1007/s11042-023- 16745-4
6. Mythili J, Surendhar T, Suryaprakash P, Kumar KS (2023) Machine Learning Tech niques for Diabetes Prediction: A Comparative Analysis. In: 2023 International Confer ence on Sustainable Computing and Smart Systems (ICSCSS). pp 177–183
7. Jyothi UP, Dabbiru M, Bonthu S, et al (2023) Comparative Analysis of Classification Methods to Predict Diabetes Mellitus on Noisy Data. In: Doriya R, Soni B, Shukla A, Gao X-Z (eds) Machine Learning, Image Processing, Network Security and Data Sci ences. Springer Nature Singapore, Singapore, pp 301–313
8. Abnoosian K, Farnoosh R, Behzadi MH (2023) Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 24:337. https://doi.org/10.1186/s12859-023-05465-z

9. Alzyoud M, Alazaidah R, Aljaidi M, et al (2024) Diagnosing diabetes mellitus using machine learning techniques. International Journal of Data and Network Science 8:179– 188. https://doi.org/10.5267/j.ijdns.2023.10.006
10. Gupta H, Varshney H, Sharma TK, et al (2022) Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex & Intel ligent Systems 8:3073–3087. https://doi.org/10.1007/s40747-021-00398-7
11. Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54:1937–1967. https://doi.org/10.1007/s10462-020-09896-5
12. Biau G, Cadre B, Rouvière L (2019) Accelerated gradient boosting. Mach Learn 108:971–992. https://doi.org/10.1007/s10994-019-05787-1
13. Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227. https://doi.org/10.1007/s11749-016-0481-7
14. Shaheed K, Szczuko P, Abbas Q, et al (2023) Computer-Aided Diagnosis of COVID 19 from Chest X-ray Images Using Hybrid-Features and Random Forest Classifier. Healthcare (Switzerland) 11:. https://doi.org/10.3390/healthcare11060837
15. Dataset Link: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-da taset